International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 4 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195435

Comparative Study of Microarray Based Disease Prediction - A Survey

T. Sneka[1*], K. Palanivel[2]

[1]Research Scholar, Department of Computer Science, A.V.C. College (Autonomous) Mayiladuthurai, Tamil Nadu, India

[2]Associate Professor, Department of Computer Science, A.V.C. College (Autonomous) Mayiladuthurai, Tamil Nadu, India

Corresponding Author : palani.avcc@gmail.com

ABSTRACT

Recognition of genetic expression becomes an important issue for research while diagnosing genetic diseases. Microarrays are considered as the representation for identifying gene behaviors that may help in detection process. Hence, it is used in analyzing samples that may be normal or affected, also in diagnosing various gene-based diseases. Various clustering and classification techniques were used to face the challenges in handling microarray. High dimensional data is one of the major issues caused while handling microarray. Also because of this issue, possibilities of redundant, irrelevant and noisy data may occur. To solve this problem feature selection process which optimally extracts the features is introduced in clustering in classification techniques. This survey observes some various techniques of classification, clustering of genes and feature selection methods such as supervised, unsupervised and semi-supervised methods. To determine the suitable semi-supervised algorithm that combines and analyze for detecting new or difficult mutated disease. This survey shows that how semi-supervised approach evolves and outperforms the existing algorithms.

Keywords -Micro-Array, Gene Clustering, Classification, Semi-Supervised Approach, Features Selection

I. INTRODUCTION

DNA (Deoxyribo Nucleic acid) is small spots that fixed in a glass slide called microarray. A gene may contain more number of DNA molecules. Under two different conditions say condition A and condition B same set of genes are compared. Here, clustering technique plays a vital role in identifying how gene functions are modeled suitable for corresponding cell structure. Co-expressed genes i.e., similar expression pattern of genes that are taken from two different cell can be clustered together with similar cellular function. A good clustering algorithm should be able to identify "true" number of clusters with the given data without requiring pre-determined number of clusters. Also, microarray experiments involve complicated procedures and huge amount of noise. The basic layout of the symbols can be shown in figure 1.

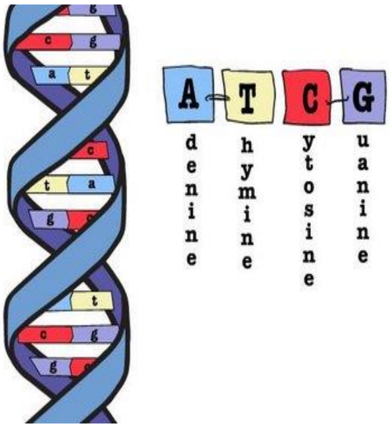

Figure 1. Gene Symbols

CSEIT195333 | Received : 15 July 2019 | Accepted : 08 August 2019 | July-August -2019 [5 (4) : 189-197]

Classification of various diseases by the use of clustered features is done by adopting various classification techniques. Based on the result of classification performance evaluation and accuracy of each classifier that is adopted to predict the disease is calculated. In this survey we take seven different microarray based disease prediction works and discuss how semi-supervised learning method provides better result. Also this survey point out the techniques that are incorporated by those seven research works which further helps to identify the advantage and disadvantage of the each work. The main contributions of this survey are as follows:

- To discuss various clustering, feature selection and classification techniques adopted for microarray based disease prediction.
- To determine how supervised, unsupervised and semi-supervised methods contribute in prediction of disease.
- Also, how semi-supervised method evolving with improved accuracy.

The organization of this document is as follows. In Section 2 (**Related Works**) seven various works related to microarray problems is discussed. In Section 3 (**Outcomes of Related Works**) the advantage and disadvantages of each work is mentioned. Various machine learning methods used in microarray based disease prediction is discussed in Section 4(**Machine Learning Approaches**) and finally, Section 5(**Conclusion)** that sums up the preferable research approaches for microarray based disease prediction and suggest some future scopes.

II. RELATED WORKS

Mohammed Aledhari, et.al, [1] designed a data minimization algorithm to transfer big genomic datasets in an expedient, secure way to allow scientists to share their data and analysis. This work used the Hyper Text Transfer Protocol (HTTP) as a baseline protocol to compare and assess implementation results of transferring big genomic datasets. The goals of our data minimization algorithm are as follows:

- reduce the size of data to be transferred between a server and a client
- Secure and protect the privacy of the data from unauthorized access due to attacks or data breach, such as Man-in-the-middle (MITM) attack.

This work proposed data minimization strategy that exploits alphabet limitation of DNA sequences. It uses CNN (convolutional Neural Network) which is deep learning-based algorithm minimizes code word used for transmitting big genome DNA datasets with different time slot.

Yvan Saeys, et.al, [2] focused on the application of feature selection techniques. In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables, hence, offering the advantage of interpretability by a domain expert. While feature selection applied to both supervised and unsupervised learning, the focus gets here on the problem of supervised learning (classification), where the class labels are known before hand. The interesting topic of feature selection for unsupervised learning (clustering) is a more complex issue, and research into this field is recently getting more attention in several communities.

Sebastian Maldonado and Richard Weber [3] proposed method is that different runs of the algorithm may select different features. This is due to the random data split in each iteration. An unfortunate split of the data set may also remove an important feature, affecting thus negatively the classifier's performance. To avoid this situation, performing 3-4 runs of the algorithm is recommended, comparing the eliminated features and removing them only if they have been discarded in more than one run. Also performance of the algorithm is checked by analyzing the number of errors and identifying incorrectly removed features, improving the method's effectiveness. Empirically proved the method's robustness regarding feature selection by verifying that most of the time the same features are selected in different runs providing high classifier performance. For example, after running the proposed method five times on the WBC (White Blood Cells) data set, 9 from the original 30 features have been selected five times. In order to obtain the features in terms of relevance, using a fast filter method for example, before running the algorithm, in order to decide which variable to remove in case of equal number of validation errors. This point is particularly important in high-dimensional data sets with a small number of observations.

Pablo Bermejo, et.al, [4] focused on the framework described above (supervised classification and the use of accuracy as performance measure), but in addition adding the constraint of dealing with high-dimensional datasets, that is, this work interested in carrying out FSS (Feature Subset Selection) over datasets having thousands of variables (microarrays, text-mining, etc.). The goal is to develop an algorithm able to perform FSS efficiently in high-dimensional datasets, and so we aim to maintain (or improve) the performance (accuracy and compactness) of previous approaches in the literature for dealing with this type of datasets. Feature subset selection is a key problem in the data-mining classification task that helps to obtain more compact and understandable models without degrading (or even improving) their performance. This work focus on FSS in high-dimensional datasets, that is, with a very large number of predictive attributes. In this case, standard sophisticated wrapper algorithms cannot be applied because of their complexity, and computationally lighter filter-wrapper algorithms have recently been proposed. This work proposed a stochastic algorithm based on the GRASP (Greedy Randomized Adaptive Search Procedure) meta-heuristic, with the main goal of speeding up the feature subset selection process, basically by reducing the number of wrapper evaluations to carry out. GRASP is a multi-start constructive method which constructs a solution in its first stage, and then runs an improving stage over that solution.

B. Chandra [5] implemented Feature selection methods can be categorized into filter, wrapper, and embedded or hybrid. Filter approach selects features without involving any learning algorithm. The filter model relies on general characteristics of the data to evaluate and select feature subsets. The wrapper approach selects feature subset based on the classifier and ranks feature subset using predictive accuracy or cluster goodness. It is more computationally expensive than the filter model. Most of the algorithms mentioned above require computationally expensive search strategy to find an optimal feature subset. However, there is a drawback in the ERGS (Energy Remaining Greedy Scheduling) algorithm that the Inter feature dependence over all the classes is not taken into account. Hence a new feature selection algorithm has been proposed algorithm has been proposed in this paper to remove the drawbacks to improve the performance Effective range based feature selection. Inter feature overlap is found using the sum (over all the classes) of overlap between the features having higher weight and all other features. The features having higher weights and least sum of overlap between the features over all the classes is included in the selected feature set. Modified weights for the features are computed by taking the difference previous weights with the Inter class overlap.

Lu Huijuan, et.al, [6] proposed a hybrid feature selection method combining MIM (Mutual information maximization) and AGA (Adaptive genetic algorithm) and name it as MIMAGA-Selection algorithm. The MIMAGA-Selection algorithm effectively reduces the dimension of the original gene expression datasets and removes the redundancies of the data. For datasets with the number of genes up to 20,000 the MIMAGA-Selection algorithm is always capable to reduce the gene number to below 300 with reasonably high classification accuracies. The classification accuracy rates comparison with other existing feature selection algorithms shows the effectiveness of the MIMAGA-Selection algorithm. Four different classifiers, namely BP (Back Propagation), SVM (Support Vector Machine), ELM (Extreme Learning Machine) and RELM (Regression Extreme Learning Machine) are applied to the reduced dataset. The lowest classification accuracy is around 80% which is still in the acceptable region. Hybrid approaches combine two or more well-studied algorithms to form a new strategy to solve a particular problem. The hybrid approach usually capitalizes on the advantages from the sub-algorithms and therefore is more robust comparing with traditional approaches. Known hybrid approaches include ensemble classifiers and hybrid feature selection methods.

Juan Ramos, et.al, [7] can be generically defined the process of extracting gene subsets whose expression level values are representative of a particular target feature, i.e., clinical or biological annotation. GS (Gene Selection) is a very active research area in the analysis of gene expression microarray, which is contributing to the development of the field as a result of involved data mining and machine learning techniques. Particularly, GS from microarrays is addressed to identify/discover those genes which are expressed differentially according to a determined target disease (namely informative genes). GS methods have been divided into the following four categories: filters, wrappers, embedded and ensemble. Filter methods have been directed to discriminate or filter features/ genes based on the intrinsic properties of the dataset. They do this by estimating their relevance scores to state a cut-off schema where an upper/lower bound is imposed to choose features with the best scores. Wrapper methods use a classifier to find the most discriminate feature subset by minimizing an error prediction function. Embedded methods are similar to wrapper but additionally they interact with the learning model, which reduces the runtime taken by wrapper methods.

III. OUTCOMES OF RELATED WORK

Table 1 provides comparative analysis of reviewed papers

S.NO	RESEARCH WORKS	PROS	CONS
1	A Deep Learning-Based Data Minimization Algorithm for Fast and Secure Transfer of Big Genomic Dataset (Mohammed Aledhari, et.al, 2018)	Minimize the gene datasets and secured transmission of genome DNA datasets.	There is no approach for disease prediction
2	A review of feature selection techniques in bioinformatics (YvanSaeys, et.al, 2007)	Dimensionality reduction	Only Univariate genes are analyzed

3	A wrapper method for feature selection using Support Vector Machines(Sebastian Maldonado and Richard Weber, 2009)	Limited gene features are analyzed	Unbalanced model construction
4	A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets (Pablo Bermejo, 2011)	Analyzed high-dimensional datasets	Provide computational complexity
5	An Efficient Feature Selection Technique for Gene Expression Data (B Chandra, 2018)	A measure to remove redundancy and to select important features	Features may be overlapped at the time of calculating gene sequence
6	A hybrid feature selection algorithm for gene expression data classification (Lu Huijuan, et.al, 2017)	Eliminate the redundant samples and reduce the dimension of the gene expression data	Provide the lowest classification accuracy
7	An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray (Juan Ramos, et.al, 2017)	Discover differentially expressed genes for a particular target annotation	Outliers may be occurred at the time of grouping similar genes

IV. MACHINE LEARNING APPROACHES

Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene-based clustering presents several new challenges and is still an open problem. First, cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis. Therefore, algorithms for gene-based clustering should be able to effectively handle this situation. Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters (e.g., which clusters are more close to each other, and which clusters are remote from each other), and the relationship between the genes within the same cluster (e.g., which gene can be considered as the representative of the cluster and which genes are at the boundary area of the cluster). A clustering algorithm, which can not only partition the data set but also provide some graphical representation of the cluster structure, would be more favored by the biologists.

A. Supervised learning algorithm:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

The steps are defined in supervised learning can be listed as follows:

- Determine the type of training examples. Before doing anything else, the user should decide what kind of data is to be used as a training set. In case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.
- Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.
- Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees.
- Complete the design. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.
- Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

The mostly used supervised learning algorithms are:

- Support Vector Machines
- linear regression
- logistic regression
- naive Bayes
- linear discriminant analysis
- decision trees

The supervised learning algorithm provided computational complexity because of large number of datasets for trained and also support only case sensitive query.

B. Unsupervised learning algorithm:

Unsupervised learning is a branch of machine learning that learns from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. Alternatives include supervised learning and reinforcement learning. A central application of unsupervised learning is in the field of density estimation in statistics, though unsupervised learning encompasses many other domains involving summarizing and explaining data features. Compared to supervised learning where training data is labeled with the appropriate classifications, models using unsupervised learning must learn relationships between elements in a data set and classify the raw data without "help." This hunt for relationships can take many different algorithmic forms, but all models have the same goal of mimicking human logic by searching for indirect hidden structures, patterns or features to analyze new data.

The mostly used unsupervised learning algorithms are:

- Principal component analysis
- Independent component analysis
- Non-negative matrix factorization
- Singular value decomposition

The classical example of unsupervised learning in the study of neural networks is Donald Hebb's principle, that is, neurons that fire together wire together. In Hebbian learning, the connection is reinforced irrespective of an error, but is exclusively a function of the coincidence between action potentials between the two neurons. A similar version that modifies synaptic weights takes into account the time between the action potentials (spike-timing-dependent plasticity or STDP). Hebbian Learning has been hypothesized to underlie a range of cognitive functions, such as pattern recognition and experiential learning. One of the statistical approaches for unsupervised learning is the method of moments. In the method of moments, the unknown parameters (of interest) in the model are related to the moments of one or more random variables, and thus, these unknown parameters can be estimated given the moments. The moments are usually estimated from samples empirically. The basic moments are first and second order moments. For a random vector, the first order moment is the mean vector, and the second order moment is the co-variance matrix (when the mean is zero). Higher order moments are usually represented using tensors which are the generalization of matrices to higher orders as multi-dimensional arrays. In particular, the method of moments is shown to be effective in learning the parameters of latent variable models. Latent variable models are statistical models where in addition to the observed variables, a set of latent variables also exist which are not observed. A highly practical example of latent variable models in machine learning is the topic modeling which is a statistical model for generating the words (observed variables) in the document based on the topic (latent variable) of the document. In the topic modeling, the words in the document are generated according to different statistical parameters when the topic of the document is changed. It is shown that method of moments (tensor decomposition techniques) consistently recovers the parameters of a large class of latent variable models under some assumptions. The major disadvantage of the unsupervised algorithm can be provided irrelevant results in gene disease prediction.

C. Semi-supervised Learning algorithm

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings. In this module implement K nearest neighbor algorithm to classify the various types of diseases from gene expression. Classification is done with the help of KNN classifier. In the recent years, KNN classifiers have established excellent performance in a variety of pattern recognition troubles. The input space is planned into a high dimensional feature space. Then, the hyper plane that exploits the margin of separation between classes is constructed. The points that lie closest to the decision surface are called support vectors directly involves its location. When the classes are non-separable, the optimal hyper plane is the one that minimizes the probability of classification error. Initially input image is formulated in feature vectors. Then these feature vectors mapped with the help of kernel function in the feature space. And finally division is computed in the feature space to separate out the classes for training data. A global hyper plane is required by the KNN in order to divide both the program of examples in training set and avoid over fitting. This phenomenon of KNN is higher in comparison to other machine learning techniques which are based on artificial intelligence. Here the important feature for the classification is the width of the vessels. With the help of KNN classifier we can easily separate out the vessels into arteries and veins. The KNNs demonstrate various attractive features such as good generalization ability compared to other classifiers. Indeed, there are relatively few free parameters to adjust and it is not required to find the architecture experimentally. The KNNs algorithm separates the classes of input patterns with the maximal margin hyper plane. This hyper plane is constructed as:

$$f(x) = \langle w, x \rangle + b$$

Where x is the feature vector, w is the vector that is perpendicular to the hyper plane and $b||w||^{-1}$ specifies the offset from the beginning of the coordinate system. To benefit from non-linear decision boundaries the separation is performed in a feature space F, which is introduced by a nonlinear mapping φ the input patterns. This mapping is defined as follows:

$$\langle \varphi\ (x_1), \varphi(x_2) \rangle = K(x_1, x_2)\ \forall (x_1, x_2) \in X$$

for some kernel function K (·, ·). The kernel function represents the non-linear transformation of the original feature space into the F.

V. CONCLUSION

Microarray is an important tool for cancer classification at the molecular level. It monitors the expression levels of large number of genes in parallel. With large amount of expression data obtained through microarray experiments, suitable statistical and machine learning methods are needed to search for genes that are relevant to the identification of different types of disease tissues. In this paper, different type learning algorithms to classify the diseases based on gene datasets is discussed. Also, it includes supervised, unsupervised and semi-supervised learning approaches. This survey concludes that semi-supervised learning algorithm is highly preferable for improved and accurate result in disease prediction. In future weighted classifiers and fuzzy as well as rough set theory can be adopted for avoiding vagueness in genetic data.

VI. REFERENCES

[1]. Mohammed Aledhari, Marianne Di Pierro, Mohamed Hefeida, Fahad Saeed. "A deep learning-based data minimization algorithm for fast and secure transfer of big genomic datasets". IEEE Transactions on Big Data . pp. 1-1.2018.

[2]. Yvan Saeys, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics". Bioinformatics,.Vol. 23.19, pp. 2507-2517, 2007.

[3]. Sebastian Maldonado, and Richard Weber. "A wrapper method for feature selection using support vector machines". Information Sciences .vol.179.13, pp. 2208-2217. 2009.

[4]. Pablo Bermejo, Jose A. Gamez, and Jose M. Puerta. "A GRASP algorithm for fast hybrid

(filter-wrapper) feature subset selection in high-dimensional datasets". Pattern Recognition Letters. vol. 32.5 ,pp.701-711.2011.

[5]. B. Chandra. "An efficient feature selection technique for gene expression data". IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). 2018.

[6]. Lu Huijuan, et.al, "A hybrid feature selection algorithm for gene expression data classification". Neurocomputing.vol. 256, pp. 56-62.2017.

[7]. Juan Ramos, et.al, "An agent-based clustering approach for gene selection in gene expression microarray". Interdisciplinary Sciences: Computational Life Sciences .vol.9.1, pp. 1-13.2017.

Cite this article as :

T. Sneka1, K. Palanivel, "Comparative Study of Microarray Based Disease Prediction - A Survey", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 189-197, July-August 2019. Available at doi : https://doi.org/10.32628/CSEIT195435
Journal URL : http://ijsrcseit.com/CSEIT195435